BIOSTATS 640 Intermediate Biostatistics
Fall 2023
Examination III

Unit 6 – Analysis of Variance
Unit 7 - Logistic Regression
Due: Wednesday December 13, 2023

last date for submission for credit (-10 points) due: Friday December 15, 2023

Before you begin:

This is a "take-home" exam. You are welcome to: 1) use any reference materials you wish; 2) use the computer as you wish (including online apps); and 3) contact me with questions.

However, you MUST work this exam **by yourself,** and you may **not** consult with anyone (except me and that is fine…)

Please:

__**1. Name your file as instructed below; and**
__**2. Complete the signature page**

How to name your exam submission

__a) Please be sure your name is somewhere on your submission.
__b) Next, save it as a SINGLE FILE pdf (please do not submit a word file)
__c) Suggestion: Use the following naming convention: lastname_exam3.pdf.

How to submit your exam in Canvas.

At left, click **ASSIGNMENTS** > Upcoming Assignments > *scroll down* > UPLOAD > Drag folder here or choose a file to upload > Submit Assignment

Questions?

Again, questions are always welcome; email me at **cbigelow@schoolph.umass.edu**

BIOSTATS 640 Intermediate Biostatistics
Fall 2023
Examination III

Unit 6 – Analysis of Variance
Unit 7 - Logistic Regression
Due: Wednesday December 13, 2023
last date for submission for credit (-10 points) due: Friday December 15, 2023

Signature

This is to confirm that in completing this exam, I worked independently and did not consult with anyone.

Name: _____

Date: _____

Thank you!

Before You Begin

This test has six (6) questions, with points that total = 100.

I have also provided one 5-point ***"extra point"*** question. ***This is optional!***

Extra points will be added to your overall test score, up to a maximum of 100.

1. (20 points total)

1a. (2 points)

Are you a morning person, an evening person, or neither? Does this personality trait affect how well you perform? A study was made of 100 students comprised of 16 "morning" people, 30 "evening" people and 54 "neither" people. Each student took a test of their ability to memorize at 8 am and again at 9 pm. Your analysis is of the difference defined as Y = [score at 8 am] - [score at 9 pm]. This is a situation where we want to compare the mean responses in several populations (we might have said "groups" here) using a one way anova.

i) What is the population?
ii) What is the response variable?
iii) What is the "k" in this one way anova?
iv) What are the values of the n_i?
v) What are the degrees of freedom of the analysis of variance F-test?

1b. (2 points)

According to CDC charts, a healthy BMI for an eight-year-old girl is between 13.5 and 18.3. A study was conducted of the body mass index (BMI) of eight-year old girls in 3 national surveys: NHES II, NHANES II 1980, and NHANES 2002. The following table shows selected summary statistics from this study.

Survey	Survey Sample size	Survey Sample mean	Survey Sample Standard Error of the Mean
NHES II, 1965	613	16.4	0.1
NHANES II, 1980	125	16.3	0.2
NHANES, 2002	184	18.3	0.5

Study investigators want to compare the mean responses in several populations using a one way anova.

i) What is the population?
ii) What is the response variable?
iii) What is the "k" in the anova design?
iv) What are the values of the n_i?
v) What are the degrees of freedom of the anova F-test?

1c. (2 points)

The Quebec (Canada) Cardiovascular Study is investigating the relationship between weight and triglyceride (mmol/l). It recruited men aged 34 to 64 at random from towns in the Quebec City metropolitan area. At initial screening, 1824 invitees met the criteria (no diabetes, free of heart disease, and so on). Eligible and consenting participants were classified into 3 groups defined by weight: normal weight (n=719, mean triglyceride level 1.5 mmol/l), overweight (n=885, mean triglyceride level 1.7 mmol/l) and obese (n=220, mean triglyceride level 1.9 mmol/l).

Researchers want to compare the mean triglyceride levels across groups using a one way anova.

i) What is the population?
ii) What is the response variable?
iii) What is the "k" in this anova design?
iv) What are the values of the n_i?
v) What are the degrees of freedom of the anova F-test?

1d. (2 points)

The following ANOVA table is only partially completed. Complete the table by filling in the **??**.

Source	DF	Sum of Squares	Mean Square	F-statistic	p-value
Between groups	3	??	45	??	??
Within groups	12	337	??		
Total, corrected	??	472			

1e. (2 points)

The following ANOVA table is also only partially completed.

Source	DF	Sum of Squares	Mean Square	F-statistic	p-value
Between groups	??	258	??	??	??
Within groups	26	??	??		
Total, corrected	29	898			

i) Complete the table by filling in the **??**.
ii) How many groups were there in the study?
iii) How many total observations were there in the study?

1f. (5 points)

In an analysis of the General Social Survey (GSS), multiple predictor normal theory linear regression was performed. Specifically, y=number of hours per day spent watching TV was regressed on x_1 = gender (1=male, 0=female), x_2 =political affiliation is Republican (1=yes, 0=no), x_3= political affiliation is Democrat (1=yes, 0=no), and x4=political affiliation is Libertarian (1=yes, 0=no). The referent political affiliation is Independent. The following estimated prediction equation was obtained:

$$\hat{y} = 2.4 + 0.2x_1 + 0.5x_2 + 0.8x_3 - 0.1x_4$$

i) **(1 point)** Interpret the gender effect.

ii) **(1 point)** Interpret the coefficient of x_2

iii) **(3 points)** State the equivalent analysis of variance model. Use ***deviation from means parameterization.*** Define all terms and constraints on the parameters.

1g. (5 points)

A multiple predictor normal theory linear regression analysis of college faculty salaries included an indicator variable for sex at birth (1=male, 0=female) and an indicator for Non-White race (1=Non-White, 0 = White). For annual income measured in thousands of dollars, the estimated regression coefficients were 0.8 for sex at birth and 0.6 for race.

i) **(1 point)**
Interpret the coefficient for sex at birth.

ii) **(2 points)**
At particular settings of the other predictors, the estimated mean salary for white females was 60.2 thousand dollars. Using the estimated coefficients, find the estimated means for White males:

iii) **(2 points)**
Find the estimated means for Non-White females.

2. (20 points total)

2a. (4 points)

Multiple choice. Pick ONE. The purpose of analysis of variance is to compare

__ a. the variances of several populations.
__ b. the proportions of successes in several populations.
__ c. the means of several populations.

2b. (4 points)

Multiple choice. Pick ONE. A study of the effects of smoking classifies subjects as nonsmokers, moderate smokers, or heavy smokers. The investigators interview a sample of 200 persons in each group. Among the questions is "how many hours do you sleep on a typical night?" The degrees of freedom for the analysis of variance F statistic for comparing mean hours of sleep are

__a. 2 and 197
__b. 2 and 597
__c. 3 and 597

The following setting pertains to Question 2 parts c, d, and e.

A dental study evaluated the effect of tooth etch time on resin bonding strength. A total of 78 undamaged, recently extracted first molars (baby teeth) were randomly assigned to be etched with phosphoric acid gel for either 15, 30, or 60 seconds. Composite resin cylinders of identical size were then bonded to the tooth enamel. The researchers examined the bond strength after 24 hours by finding the failure load (in megapascal) for each bond. Here are the summary data and ANOVA table for this experiment.

Etch time	n	$\overline{x}$	s
15 seconds	26	4.49	2.28
30 seconds	26	6.98	3.15
60 seconds	26	8.48	4.17

Source	DF	Sum of Squares	Mean Square	F-statistic	p-value
Etch time	2	211.208	105.604	9.745	0.0002
Error	75	812.745	10.837		
Total, corrected	77	1023.953			

2c. (4 points)

Multiple choice. Pick ONE. The most striking conclusion from the numerical summaries for the three etch times is that

__a. there appears to be little difference among the etch times.
__b. on average, failure load increases with etch time
__c. on average, failure load decreases with etch time.

2d. (4 points)
Multiple choice. Pick ONE. The conclusion of the analysis of variance test is that

__a. there is quite strong evidence (p=.0002) that the mean failure loads are not the same in all three conditions.
__b. there is quite strong evidence (p=.0002)that the mean failure load is much lower for 15 seconds etch time than in any other two etch times.
__c. the data give no evidence (p=.0002) to suggest that mean failure load differ for the three etch times.

2e. (4 points)
Multiple choice. Pick ONE. If we used a series of two-sample t procedures to compare the three conditions, we would have to give three 95% confidence intervals to compare all three pairs of etch times. The weakness of doing this is that

__a. we won't be 95% confident that all 3 intervals cover the true differences in population means.
__b. the conclusions from the three intervals might not agree.
__c. the conditions for two-sample t inference are not met for all 3 pairs of etch times.

3. (20 points total)

In a logistic regression analysis of the likelihood (π) of mortality that considered several variables, a one predictor model was fit to malnutrition (malnut) coded 1=malnutrition, 0=NO malnutrition. The following estimated logit prediction equation was obtained:

$$\hat{\text{logit}}[\hat{\pi}] = -1.8563 + 1.210[\text{malnut}]$$

The 2x2 table associated with these data is the following

		Mortality		
		1 = Dead	0=Alive	
malnut	1=Malnourished	11	21	32
	0=NOT malnourished	10	64	74
		21	85	106

3a. (8 points)
Verify that the regression coefficient (beta) for malnut in the logistic regression model is the natural logarithm of the odds ratio for malnut in the 2x2 table. Show all work.

3b. (3 points)
Using the logistic regression model, what is the ***formula*** for the predicted probability of death for a person who is malnourished?

3c. (3 points)
Using the logistic regression model, what is the ***numeric value*** of the predicted probability of death for a person who is malnourished?

3d. (3 points)
Using the 2x2 table, what is the ***formula*** for the empirical estimate of the probability of death for a person who is malnourished? ***Hint – the empirical estimate is simply the observed proportion.*** What is its calculated numeric value?

3e. (3 points)
Using the 2x2 table, what is the ***numeric value*** of the empirical estimate of the probability of death for a person who is malnourished?

4. (10 points total)

A multiple predictor logistic regression was performed to identify the correlates of home ownership Y (1=yes, 0=no) among young married couples. The following table is a partial display of the results of fitting a 7-predictor model.

Variable (predictor)	**Regression Coefficient** $\hat{\beta}$	**Standard Error** $\hat{se}(\hat{\beta})$
Intercept (constant)	-2.870	-
Husband earnings, in $10,000's	0.569	0.088
Wife earnings, in $10,000's	0.306	0.140
Number of years married	-0.039	0.042
Working wife in 2 years (1=yes, 0=no)	0.373	0.283
Number of children	0.220	0.101
Add child in 2 years (1=yes, 0=no)	0.271	0.140
Parents own their homes (1=yes, 0=no)	0.387	0.176

4a. (5 points)

Using the regression coefficients and standard errors provided, carry out appropriate Wald tests to identify which variables are statistically significantly associated with home ownership. In 1-4 sentences, interpret your findings.

4b. (5 points) Fill in the blanks

Adjusting for the other explanatory variables in the model, each additional child is estimated to multiply the odds of homeownership by OR = BLANK , that is, the estimated odds increase by BLANK.

Whereas an increase of $10,000 in the husband's earnings is estimated to multiply the odds of owning a home by OR = BLANK, an increase of $10,000 in the wife's earnings is estimated to multiply the odds of owning a home by OR = BLANK, after adjustment for the other explanatory variables in the model.

5. (20 points total)

A logistic regression analysis of likelihood (π) of mortality considered several variables: shock (SHOCK: coded 1=shock, 0=NO shock), malnutrition (MALNUT; coded 1=malnourished, 0 = NOT malnourished), alcoholism (ALC: coded 1=alcoholic 0=NOT alcoholic), age (AGE: continuous), and bowel infarction (INFARCT: coded 1=infarction, 0=NO infarction). The following fitted model was obtained:

$$\widehat{\text{logit}}[\hat{\pi}] = -9.754 + 3.674[\text{SHOCK}] + 1.217[\text{MALNUT}] + 3.355[\text{ALC}] + 0.09215[\text{AGE}] + 2.798[\text{INFARCT}]$$

5a. (8 points total)
What is the estimated probability of death for a 60 year old ***malnourished*** patient with no evidence of shock, but with symptoms of alcoholism and prior bowel infarction? In developing your answer:

5a. i. (4 points)
Write out the formula you use

5a. ii. (4 points)
Provide the numeric value of the estimate.

5b. (6 points total)
Write out the expression for the predicted logit of mortality as a function of age for the sub-population for whom MALNUT=0, ALC=1, SHOCK=0, and INFARCT=1.

5c. (6 points total)
Using your answer to question 5b as a start, by any means you like, construct a plot of the estimated probability of death versus age for a person who is NOT malnourished, has symptoms of alcoholism, and has a bowel infarction. ***A hand drawn plot is just fine!*** In 1-2 sentences, interpret your plot.

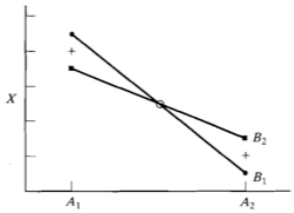
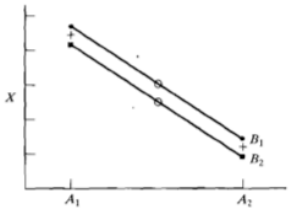
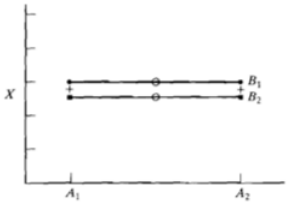
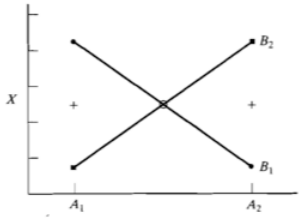
6. (10 points total)

One of the learning objectives of BIOSTATS 640 is to develop an understanding of effect modification. A term for this in analysis of variance parlance is "interaction". A two way factorial design with replicate observations for each combination of factor I and II permits the discovery of "interaction". We can get a visual sense of this by constructing a plot of the group means. Consider a two way factorial design with factor I at two levels designated "A_1" and "A_2" and factor II at two levels, designated "B_1" and "B_2".

For each of the following three scenarios of main effects and interaction, provide the following:

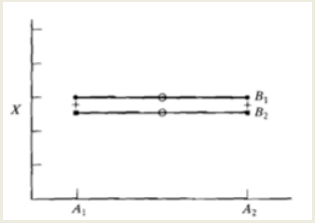
(i) Identify the correct graphical summarization of the means.

(ii) Write down the correct model using notation μ, α_i, β_j, $(\alpha\beta)_{ij}$, and σ_e^2 as appropriate.

Picture 1	Picture 2	Picture 3	Picture 4
			

I've done one for you as an example

No effect of factor I, small effect of factor II, and no interaction.

(i) **Picture 3** (ii) Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$	

6a. (3 points)
Large effect of factor I, small effect of factor II, and no interaction.

(i) Picture: (ii) Model:	

6b. (3 points)
No effect of factor I, no effect of factor II, and interaction.

(i) Picture: (ii) Model:	

6c. (4 points)
Large effect of factor I, no effect of factor II, and interaction.

(i) Picture: (ii) Model:	

Optional Extra Points

Points earned here will be added to your score up to a maximum total = 100.
Sorry, no partial credit on these.

(5 points total)

Complete the following one way of analysis of variance table, by filling in the **??.**

Source	Degrees of Freedom	Sum of Squares	Mean Square	F
Between	??	360.00	??	??
Within	15	450.00	??	
Total, corrected	19	??		